



## U-statistics in stochastic geometry

Raphaël Lachièze-Rey, Matthias Reitzner

### ► To cite this version:

Raphaël Lachièze-Rey, Matthias Reitzner. U-statistics in stochastic geometry. G. Peccati and M. Reitzner. Stochastic Analysis for Poisson Point Processes: Malliavin Calculus, Wiener-Itô Chaos Expansions and Stochastic Geometry, 7 (1), Springer International Publishing, 2016, Bocconi & Springer Series, 978-3-319-05232-8. hal-01120915v2

**HAL Id: hal-01120915**

**<https://hal.science/hal-01120915v2>**

Submitted on 25 Mar 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# U-statistics in stochastic geometry

Raphaël Lachièze-Rey and Matthias Reitzner

**Abstract** A U-statistic of order  $k$  with kernel  $f : \mathbb{X}^k \rightarrow \mathbb{R}^d$  over a Poisson process is defined in [25] as

$$\sum_{x_1, \dots, x_k \in \eta_{\neq}^k} f(x_1, \dots, x_k)$$

under appropriate integrability assumptions on  $f$ . U-statistics play an important role in stochastic geometry since many interesting functionals can be written as U-statistics, like intrinsic volumes of intersection processes, characteristics of random geometric graphs, volumes of random simplices, and many others, see for instance [15, 18, 25]. It turns out that the Wiener-Ito chaos expansion of a U-statistic is finite and thus Malliavin calculus is a particularly suitable method. Variance estimates, the approximation of the covariance structure and limit theorems which have been out of reach for many years can be derived. In this chapter we state the fundamental properties of U-statistics and investigate moment formulae. The main object of the chapter is to introduce the available limit theorems.

## 1 U-statistics and decompositions

### 1.1 Definition

Let  $\mathbb{X}$  be a Polish space,  $k \geq 1$ , and  $f : \mathbb{X}^k \rightarrow \mathbb{R}$  be a measurable function. The  $U$ -statistic of order  $k$  with kernel  $f$  over a configuration  $\eta \in \mathbf{N}_s(\mathbb{X})$  is 0 if  $\eta$  has strictly less than  $k$  points and the formal sum

---

Name of First Author  
Name, Address of Institute, e-mail: name@email.address  
Name of Second Author  
Name, Address of Institute e-mail: name@email.address

$$U(f; \eta) = \sum_{\mathbf{x}_k \in \eta_{\neq}^k} f(\mathbf{x}_k)$$

otherwise, where  $\eta_{\neq}^k$  is the class of  $k$ -tuples  $\mathbf{x}_k = (x_1, \dots, x_k)$  of distinct points from  $\eta$ . Remark that since the sum is over all such  $k$ -tuples,  $f$  can be assumed to be symmetric without loss of generality.

An abundant literature deals with the asymptotic study of  $U(f; \tilde{\eta}_p)$  as  $p \rightarrow \infty$  when  $\tilde{\eta}_p$  is a binomial process, i.e. a set of  $p$  iid variables over  $\mathbb{X}$ . We are concerned here with Poisson input, i.e.  $\eta$  is a Poisson measure over  $\mathbb{X}$  which intensity is a non-atomic locally finite measure  $\mu$  on  $\mathbb{X}$ . So that the definition makes any sense, the basic assumption is that  $f \in L_s^1(\mathbb{X}^k) = L_s^1(\mathbb{X}^k; \mu^k)$ .

In the sequel of this section, let  $\mu$  be a non-atomic locally finite measure on  $(\mathbb{X}, \mathcal{X})$ ,  $\eta$  a Poisson measure with intensity  $\mu$ , and  $k \geq 1$ .

## 1.2 chaotic decomposition and multiple integrals

**Theorem 1.** *Let  $f \in L_s^1(\mathbb{X}^k)$  such that  $U(f; \eta) \in L^2(\mathbb{P})$ . We have the  $L^2$  decomposition*

$$U(f; \eta) = \sum_{n=0}^k I_n(h_n). \quad (1)$$

Here  $I_n$  is the  $n$ -th order stochastic integral over  $\eta$  defined in Chapter [17]. The functions  $h_n$  have been explicitly computed in [25, Lemma 3.3],

$$h_n(\mathbf{x}_n) = \binom{n}{k} \int_{\mathbb{X}^n} f(\mathbf{x}_n, \mathbf{x}_{k-n}) d\mu^{k-n}(\mathbf{x}_{k-n}) \quad (2)$$

for  $\mathbf{x}_n \in \mathbb{X}^n$ , and  $h_n$  is a function of  $L_s^1(\mathbb{X}^n) \cap L_s^2(\mathbb{X}^n)$ .

*Remark 1.* Somewhat counterintuitively,  $f \in L_s^1(\mathbb{X}^k) \cap L_s^2(\mathbb{X}^k)$  does not imply that  $\mathbb{E}U(f; \eta)^2 < \infty$  (see [25]), but in most examples  $f$  is bounded and has a bounded support, which makes the latter condition automatically satisfied.

As is apparent in Theorem 1, each  $U$ -statistic of order  $k$  is a finite sum of multiple integrals of order  $n \leq k$ , and it is not difficult to prove that conversely any multiple integral of order  $n \geq 1$  can be written as a finite sum of  $U$ -statistics which orders are smaller or equal to  $n$ . From a formal point of view, it is therefore equivalent to study the asymptotics of finite sums of  $U$ -statistics or of finite sums of multiple integrals.  $U$ -statistics are more likely to appear in applications, but the homogeneity of multiple integrals make them easier to deal with, and some of the Malliavin operators of  $U$ -statistics have a particularly intuitive form. Consider for instance the case where  $F = I_k(f)$  is a multiple integral of order  $k$ . The Malliavin derivative, the

Orstein-Uhlenbeck, and inverse Orstein-Uhlenbeck operators, take the following form

$$D_x F = k I_{k-1}(f(x, \cdot)), x \in \mathbb{X}, \quad L F = -k I_k(f), \quad L^{-1} F = -k^{-1} I_k(f). \quad (3)$$

For a  $U$ -statistic  $F$ , one can still derive  $D_x F, L F, L^{-1} F$  using the linearity of those operators and the decomposition (1).

The object of this section is really the study of sums of multiple integrals which order is bounded by some  $k \geq 1$ . The chaotic decomposition also yields that any  $L^2$  variable can be approximated by such a sum, allowing us in some cases to pass on limit theorems stated here to infinite sums. The following result gives the first two moments of  $U$ -statistics.

**Proposition 1.** *Let  $f \in L_s^1(\mathbb{X}^k)$ . Then  $\mathbb{E}|U(f; \eta)| < \infty$  and*

$$\mathbb{E}U(f; \eta) = \int_{\mathbb{X}^k} f(\mathbf{x}_k) d\mu^k(\mathbf{x}_k).$$

*If furthermore  $U(f; \eta) \in L^2(\mathbb{P})$ ,*

$$\text{Var}(U(f; \eta)) = \sum_{n=1}^k n! \|h_n\|^2 \quad (4)$$

*where  $h_n$  is given in Theorem 1 and  $\|h_n\|$  is the usual  $L^2(\mathbb{X}^n)$ -norm.*

*Proof.* The first statement is a direct consequence of the Slyvniack-Mecke formula, while the second stems from the orthogonality between multiple integrals  $I_n(h_n), 0 \leq n \leq k$ .

### 1.3 Hoeffding decomposition

Let  $N \geq 1, \tilde{\eta}_p = \{X_1, \dots, X_p\}$  be a family of i.i.d. variables with common distribution  $\mu$  on  $\mathbb{X}$ . Given a measurable kernel  $h$  over  $\mathbb{X}^k, k \geq 1$ , the traditional Hoeffding decomposition (see e.g. Vitale [31]) is written

$$U(h, \tilde{\eta}_p) = k! \binom{p}{k} \sigma_k^N(h) = k! \binom{p}{k} \sum_{m=0}^k \binom{k}{m} \sigma_m^p(H_m),$$

where

$$\sigma_m^p(H_m) = \frac{1}{\binom{N}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq p} H_m(X_{i_1}, \dots, X_{i_m}), \quad 0 \leq m \leq k,$$

and each kernel  $H_m$  is symmetric and *completely degenerated*, i.e.

$$\mathbb{E}H_m(x_1, \dots, x_{m-1}, X_m) = \int_{\mathbb{X}} H_m(x_1, \dots, x_{m-1}, y) d\mu(y) = 0$$

for  $\mu^{(m-1)}$ -a.e.  $x_1, \dots, x_{m-1}$ . This property implies in particular the orthogonality of the  $\sigma_m^p(H_m)$ ,  $1 \leq n \leq k$ . If  $\mu$  is a probability measure, the  $H_m$  are uniquely defined and can be expressed explicitly via an inclusion-exclusion formula

$$H_m(x_1, \dots, x_m) = \sum_{n=0}^m (-1)^{m-n} \sum_{1 \leq i_1 < \dots < i_n \leq m} \binom{k}{n}^{-1} h_n(x_{i_1}, \dots, x_{i_n}) \quad (5)$$

where  $h_n$  is defined in (2). As is clear in this last formula, this decomposition is different from (1) because in the latter multiple integration is performed with respect to the compensated measure  $\eta - \mu$ , while in  $\sigma_m^p(H_m)$  the compensation occurs in the kernel  $H_m$ .

The Hoeffding rank  $m_1$  is defined as the smallest index  $m$  such that  $\|H_m\| \neq 0$ , and we can see through (5) that it is equal to the smallest index  $n$  such that  $\|h_n\| \neq 0$ . We furthermore have  $H_{m_1} = \binom{k}{m_1}^{-1} h_{m_1}$ . As proved in [7] for binomial processes or [16] for Poisson processes, the stochastic integral of order  $m_1$  dominates the sum, and limit theorems for geometric  $U$ -statistics can then be derived by studying this term, see Section 2.1.2.

#### 1.4 Contraction operators

Let  $f \in L_s^1(\mathbb{X}^q)$ ,  $g \in L_s^1(\mathbb{X}^k)$ . If  $f$  and  $g$  satisfy the technical conditions defined in Chapter [17], one can define for  $1 \leq r \leq l \leq \min(q, k)$  their contraction function of index  $(r, l)$ , denoted  $f \star_l^r g$ . It has  $k + q - r - l$  variables as arguments, decomposed in  $(\mathbf{x}_{r-l}, \mathbf{y}_{q-r}, \mathbf{z}_{k-r})$ , where  $\mathbf{x}_{r-l} \in \mathbb{X}^{r-l}$ ,  $\mathbf{y}_{q-r} \in \mathbb{X}^{q-r}$ , and  $\mathbf{z}_{k-r} \in \mathbb{X}^{k-r}$ . We have

$$f \star_l^r g(\mathbf{x}_{r-l}, \mathbf{y}_{q-r}, \mathbf{z}_{k-r}) := \int f(\mathbf{x}_l, \mathbf{x}_{r-l}, \mathbf{y}_{q-r}) g(\mathbf{x}_l, \mathbf{x}_{r-l}, \mathbf{z}_{k-r}) d\mu^l(\mathbf{x}_l).$$

Remember that each function appearing here is symmetric, whence the order of the arguments does not matter. Contraction operators are used below to assess the distance between a stochastic integral and the normal law. See [4] for more information on contraction operators.

## 2 Rates of convergence

Let  $F$  be a  $L^2$  variable of the form

$$F = \sum_{n=0}^k I_n(h_n) \quad (6)$$

for some kernels  $h_n \in L_s^2(\mathbb{X}^n)$ ,  $n \geq 1$ . We assume that those kernels satisfy the technical conditions mentioned in Chapter [4] so that their mutual contraction kernels are well defined. This model englobes  $U$ -statistics, as outlined by Theorem 1, as well as finite sums of  $U$ -statistics and multiple integrals.

In applied situations, the set-up consists of a fixed integer  $k \geq 1$ , and, for  $t > 0$ , a family of measures  $\mu_t$  on  $\mathbb{X}$ , and a family of kernels  $h_{n,t} \in L_s^2(\mathbb{X}^n; \mu_t^n)$ ,  $1 \leq n \leq k$ . We study the random variables

$$F_t := \sum_{n=1}^k I_n(h_{n,t}), \quad (7)$$

and more precisely the existence of numbers  $a_t, b_t > 0$  and of a random variable  $V$  such that

$$\tilde{F}_t := \frac{F_t - a_t}{\sqrt{b_t}} \rightarrow V$$

in the weak topology. In all the applications,  $\mu_t$  is either of the form

- $\mu_t = t\mu$  for some reference measure  $\mu$  on the space  $\mathbb{X}$ , or
- $\mu_t = 1_{\mathbb{X}_t}\mu$  where  $\mathbb{X}_t \subset \mathbb{X}$  depends on  $t$ .

The following two settings occur in the most important applications.

If  $\eta = \eta_t$  is a Poisson point process on  $\mathbb{X} = \mathbb{R}^d$  the measure  $\mu$  will often be the Lebesgue measure  $\ell_d$ , or for  $\mathbb{X} = \mathbb{R}^d \times M$  a product measure  $\mu = \ell_d \otimes \nu$  with a probability measure  $\nu$  on a topological marks space  $(M, \mathcal{M})$ .

If  $\eta = \eta_t$  is a Poisson 'flat' process on the Grassmannian  $\mathbb{X} = \mathcal{A}_i^d$  of affine  $i$ -dimensional subspaces (flats) of  $\mathbb{R}^d$ , the intensity measure  $\mu(\cdot)$  will be a translation invariant measure on  $\mathcal{A}_i^d$ . The Poisson flat process is only observed in a compact convex window  $W \subset \mathbb{R}^d$  with interior points. Thus, we can view  $\eta_t$  as a Poisson process on the set  $[W]$  defined by

$$[W] = \left\{ h \in \mathcal{A}_i^d : h \cap W \neq \emptyset \right\}.$$

## 2.1 Central Limit theorem

Let  $F$  be of the form (6). Let  $N \sim \mathcal{N}(0, 1)$ ,  $\sigma^2 = \text{Var}(F)$ . The next result, which Wasserstein bound has been established in [15], and Kolmogorov bound in [8], gives a bound on the distance between  $F$  and  $N$  in terms of the contractions between the kernels of  $F$ .

**Theorem 2.** *Put*

$$B(F) = \max \left( \max_1 \|h_n \star_r^l h_n\|, \max_2 \|h_n \star_r^l h_m\|, \max_{n=1, \dots, k} \|h_n\|^2 \right)$$

$$B'(F) = \max(|1 - \sigma^2|, B(F), B(F)^{3/2})$$

where  $\max_1$  is over  $1 \leq r \leq n \leq k, 1 \leq l \leq r \wedge (n-1)$ ,  $\max_2$  is over  $1 \leq l \leq r \leq n < m \leq k$ . There exists a constant  $C_k > 0$  not depending on the kernels of  $F$  such that

$$d_W(F, N) \leq \sigma^{-1} C_k B(F) \quad (8)$$

$$d_K(F, N) \leq C_k B'(F). \quad (9)$$

We reproduce here the important steps of the proof for the Wasserstein bound. The main result, due to Peccati, Sole, Taqqu, Utzet [20], is a general inequality on the Wasserstein distance between a Poisson functional  $F$  with variance  $\sigma^2 > 0$  having a finite Wiener-Ito expansion and the normal law. We have

$$d_W(F, N) \leq \frac{1}{\sigma} \sqrt{\mathbb{E}[(\sigma^2 - \langle D_x F, -D_x L^{-1} F \rangle_{L^2(\mathbb{X})})^2]} \quad (10)$$

$$+ \frac{1}{\sigma^2} \int_{\mathbb{X}} \mathbb{E}[(D_x F)^2 | D_x L^{-1} F|] \mu(dx).$$

To translate those inequalities into bounds on the contraction norms, we use the multiplication formula from [20], which yields that the multiplication of multiple integrals is a linear combinations of multiple integrals. For  $k, q \geq 1, f \in L_s^2(\mathbb{X}^q), g \in L_s^2(\mathbb{X}^k)$ ,

$$I_q(f) I_k(g) = \sum_{r=0}^{q \vee k} r! \binom{q}{r} \binom{k}{r} \sum_{l=0}^r \binom{r}{l} I_{q+p-r-l}(f \tilde{\star}_r^l g), \quad (11)$$

where the symmetrized contraction kernels  $f \tilde{\star}_r^l g$  are the average of kernels  $f \star_r^l g$  over all possible permutations of the variables.

If for instance  $F = I_k(f)$  is a single multiple integral, (3) gives the value of the Malliavin operators, and a computation then yields the bound (8) with  $f_k = f; f_i = 0$  for  $i \neq k$ , see [22, Prop. 5.5]. If  $F$  is a general functional with a finite decomposition, such as a  $U$ -statistic (see (1)), Malliavin operators are computed using linearity and yield the bound (8), see the proof of Th. 3.5 in [15].

Concerning Kolmogorov distance, Schulte [28, 29] has derived a Stein bound similar to (10), but with more terms on the right-hand side (Theorem 1.1), reflecting the effect that test functions are indicator functions, more irregular than the Lipschitz functions involved in Wasserstein distance. This bound was later improved by Eichelsbacher and Thäle [8, Th. 3.1], reducing the number of additional terms. With similar computations as in the Wasserstein case, one can then prove [8, Th. 4.1] that those additional terms only add contraction norms  $\|f_i \star_i^r f_j\|^{3/2}$  at the power  $3/2$ , up to a constant, yielding the bound  $B'(F)$ .

*Remark 2.* The terms in  $B'(F)$  bounding the Kolmogorov distance are smaller than the original terms present in  $B(F)$  if the bound goes to 0, and don't change the bound magnitude or its eventual convergence to 0.

*Remark 3.* The constant  $C_k$  explodes as  $k \rightarrow \infty$ . In other papers [25], [18], similar bounds are derived in more specific cases, with a different method. The constants are more tractable and allow for instance to approximate accurately the distance from a Gaussian to an infinite series of multiple integrals by that of its truncation at some order (see for instance [28]).

**Theorem 3 (4th moment theorem).** *Assume furthermore that kernels  $h_k$  are non-negative. Then for some  $C'_k > 0$*

$$B(F) \leq C'_k \sqrt{\mathbb{E}F^4 - 3\sigma^4}.$$

- In view of (8), the convergence of the 4-th moment to that of a Gaussian therefore implies central limit, with a bound for Wasserstein distance. In this case, as noted in [8], using (9) yields a similar bound for Kolmogorov distance. The positiveness of the kernels is adapted to  $U$ -statistics with a non-negative kernel.
- It is highly remarkable that the convergence of the 4-th moment to that of the Gaussian variable is therefore sufficient for such variables to converge to the normal law. The only technical requirement is that the variables  $F_t^4$  are uniformly integrable.

*Example 1 (De Jong's theorem).* Let  $f_2$  be a non-zero degenerate symmetric kernel from  $L_s^1(\mathbb{X}^2)$ , i.e. such that

$$\int_{\mathbb{X}} f_2(x, y) \mu(dx) = 0 \text{ for } \mu\text{-a.e. } y \in \mathbb{X}.$$

This degeneracy property implies that  $U(f_2, \eta) = I(f_2; \eta)$ , we also assume that  $f_2 \in L_s^4(\mathbb{X}^2)$ . De Jong [11] derived a 4-th moment central limit theorem for binomial  $U$ -statistics of the form  $U(f_2; \tilde{\eta}_p)$ , where  $p \in \mathbb{N}$  goes to infinity and  $\tilde{\eta}_p$  is a sequence of  $p$  iid variables with law  $\mu$ . In the Poisson framework, (9) yields Berry-Essen bounds between  $F = U(f_2; \eta) = I_2(f_2; \eta)$  and  $N$ :

$$\begin{aligned} d_W(\tilde{F}, N) &\leq C_2 \frac{1}{\|f_2\|^2} b(f_2) \\ d_K(\tilde{F}, N) &\leq C_2 \frac{1}{\|f_2\|^2} \max(b(f_2), b(f_2)^{3/2}) \end{aligned}$$

where

$$b(f_2) = \max(\|f_2 \star_2^0 f_2\|, \|f_2 \star_1^1 f_2\|, \|f_2 \star_2^1 f_2\|).$$



See Eichelsbacher and Thäle [8, Th. 4.5] for details. In [21], Peccati and Thäle derive bounds on the Wasserstein distance between such a  $U$ -statistic and a target Gamma variable, also in terms of contraction operators.

### 2.1.1 Local marked $U$ -statistics

For many applications, it is useful to assume that the state space  $\mathbb{X}$  is of the form  $S \times M$  where  $S$  is a subset of  $\mathbb{R}^d$  containing the points  $t_i$  of  $\eta$ , and  $(M, \mathcal{M})$  is a *mark space*, i.e. a locally compact space endowed with some probability measure  $\nu$ . The space  $M$  contains marks  $m_i$  that will be randomly assigned to each point of the process. In this setting, assume that  $\eta_t$  has intensity measure  $\mu_t = \mathbb{1}_{\mathbb{X}_t} \ell_d \otimes \nu$  and let  $F_t \in L^2(\mathbb{P})$  be a  $U$ -statistic  $F_t = U(f; \eta_t)$ . Let the kernel  $f$  of the  $U$ -statistic be locally square integrable on  $\mathbb{X}_t = [-t^{1/d}, t^{1/d}]^d \times M$  and stationary, i.e. for  $\mu^k$ -almost all  $(\mathbf{t}_k, \mathbf{m}_k) \in \mathbb{X}_t^k, z \in \mathbb{R}^d$ ,

$$f(\mathbf{t}_k + z, \mathbf{m}_k) = f(\mathbf{t}_k, \mathbf{m}_k). \quad (12)$$

The tail behavior of the function  $f$  is fundamental regarding the limit of variables  $F_t$  as  $t \rightarrow \infty$ .

**Definition 1.** A measurable function  $f : (\mathbb{R}^d \times M)^k \rightarrow \mathbb{R}$  is rapidly decreasing if it is locally square integrable, stationary, and if it satisfies the following integrability condition: There exists a non-vanishing probability density  $\kappa$  on  $(\mathbb{R}^d)^{k-1}$  such that for  $p = 2, 4$ ,

$$A_p(f) = \int_{(\mathbb{R}^d)^{k-1} \times M^k} f(0, \mathbf{t}_{k-1}, \mathbf{m}_k)^p \kappa(\mathbf{t}_{k-1})^{1-p} d\ell_d^{k-1}(\mathbf{t}_{k-1}) d\nu^k(\mathbf{m}_k) < \infty.$$

The slight abuse of notation  $f(0, \mathbf{t}_{k-1}, \mathbf{m}_{k-1})$  means that  $\mathbf{t}_k = (0, \mathbf{t}_{k-1}) = (0, t_2, \dots, t_{k-1})$ , and  $\mathbf{m}_k = (m_1, \dots, m_k)$ .

We have in this case the following result, which is a consequence of Theorem 6.2 and Example 2.12-(ii) in [16] :

**Theorem 4.** Let  $F_t = U(f; \eta_t)$  where  $f$  is a rapidly decreasing function, and  $\mu_t = \mathbb{1}_{\mathbb{X}_t} \ell_d \otimes \nu$  with  $\mathbb{X}_t = [-t^{1/d}, t^{1/d}]^d \times M$ . Then, with  $a_t = \mathbb{E}F_t, b_t = \text{Var}(F_t)$ , we have for some  $C_1, C_2, C_3 > 0$  not depending on  $t$ ,

$$C_1 t \leq b_t \leq C_2 t \\ d_W(\tilde{F}_t, N_1) \leq C_3 t^{-1/2}.$$

*Remark 4.* Reitzner & Schulte [25] first established this result in the case where  $f$  is the indicator function of a ball of  $\mathbb{R}^d$  (any non-vanishing continuous density  $\kappa$  can be chosen in this case because  $f(0, \cdot)$  has a compact support).

*Remark 5.* A similar result holds if  $F$  is simply assumed to be a finite sum of stochastic integrals which kernels are rapidly decreasing functions, the  $U$ -statistics being a particular case.

### 2.1.2 Geometric $U$ -statistics

Coming back to the general framework, assume  $F_t = U(f; \mu_t)$  where  $f \in L^2_s(\mathbb{X}^k)$  is fixed and  $\mu_t = t\mu$  for some measure  $\mu$  on  $\mathbb{X}$ . Then  $F_t$  admits the decomposition (7) where

$$h_{n,t}(\mathbf{x}_n) = t^{k-n} \binom{k}{n} \int_{\mathbb{X}^{k-n}} f(\mathbf{x}_n, \mathbf{x}_{k-n}) d\mu^{k-n}(\mathbf{x}_{k-n}), \mathbf{x}_n \in \mathbb{X}^n.$$

One can then see that the term  $\|h_{1,t}\|$  dominates the other terms in the variance expression (4), provided this term does not vanish. In any case the important feature is the Hoeffding rank of the  $U$ -statistic

$$n_1 := \inf\{n : \|h_n\| \neq 0\},$$

because it turns out that  $I_{n_1}(h_{n_1,t})$  is the predominant term in (7), in the sense that  $F_t - I_{n_1}(h_{n_1,t}) = o(F_t)$  for the  $L^2$  norm as  $t \rightarrow \infty$ . It yields the following result (Theorem 7.3 in [16]).

**Theorem 5.** *For some  $C_1, C_2, C_3 > 0$  not depending on  $t$ ,*

$$C_1 t^{2k-n_1} \leq b_t \leq C_2 t^{2k-n_1}.$$

(i) *If  $n_1 = 1$ ,  $U(f; \mu_t)$  follows a central limit theorem and*

$$\begin{aligned} d_W(\tilde{F}_t, N) &\leq C_3 t^{-1/2}, \\ d_K(\tilde{F}_t, N) &\leq C_3 t^{-1/2}. \end{aligned}$$

(ii) *If  $n_1 > 1$ ,  $U(f; \mu_t)$  does not follow a CLT and  $\tilde{F}_t$  converges to a Gaussian chaos of order  $n_1$  (see [16, Theorem 7.3-2]).*

For a deeper understanding we refer to the proof of Theorem 6.

*Remark 6.* Point (i) first appears in [25].

*Remark 7.* Point (ii) crucially uses the results of Dynkin & Mandelbaum [7].

*Remark 8.* The speed of convergence to the Gaussian chaos in (ii) is studied by Peccati and Thäle [21] in case the limit is a Gamma distributed random variable.

### 2.1.3 Regimes classification

The crucial difference in Theorems 4 and 5 is the area of influence of a given point  $x \in \eta_t$ . In the case of a local  $U$ -statistic, a typical point  $x \in \eta_t$  interacts with a stochastically bounded number of neighbors, that are more likely near in view of Assumption 12. The situation is different for a geometric  $U$ -statistic, where a point potentially interacts with any other point, regardless of the distance. Both these regimes can be seen as two particular cases of a continuum.

Let  $\alpha_t > 0$  be a *scaling factor*,  $\mathbb{X}_t = [-t^{1/d}, t^{1/d}]^d \times M$ ,  $\mu_t = \mathbb{1}_{\mathbb{X}_t} \ell_d \otimes \nu$ , and  $F_t = U(f_t; \eta_t)$ , where  $f_t$  is obtained by rescaling a rapidly decreasing function  $f$ :

$$f_t(\mathbf{x}_k) = f(\alpha_t \mathbf{x}_k), \mathbf{x}_k \in \mathbb{X}_t^k. \quad (13)$$

Say that  $f$  has non-degenerate projections if the functions

$$f_n(\mathbf{x}_n) = \int_{\mathbb{X}_t^{k-n}} f(\mathbf{x}_n, \mathbf{x}_{k-n}) d\mu_t^{k-n}, \mathbf{x}_n \in \mathbb{X}^n,$$

well defined in virtue of (12), are not  $\mu$ -a.e. equal to 0. It is trivially the case if for instance  $f \neq 0$  and  $f \geq 0$   $\mu$ -a.e.. Concerning notations, every spatial transformation of a point  $x = (t, m) \in \mathbb{R}^d \times M$ , such as translation, rotation, or multiplication by a scalar number, is only applied to the spatial component  $t$ .

Subsequently, any spatial transformation applied to a  $k$ -tuple of points  $\mathbf{x}_k = (x_1, \dots, x_k)$  is applied to the spatial components of the  $x_i$ 's. The quantity  $v_t = \alpha_t^{-d}$  is relevant because it gives the magnitude of the number of points interacting with a typical point  $x$ . The case  $v_t = \alpha_t = 1$  is that of local  $U$ -statistic. If  $v_t = t$  is roughly the volume of  $\mathbb{X}_t$ , it corresponds to geometric  $U$ -statistics. In this case it is useless to assume that  $f$  is rapidly decreasing, as only the behavior over  $\mathbb{X}_1$  is relevant for the problem.

**Theorem 6.** *Assume that  $f_t$  is of the form (13), where  $f$  is a rapidly decreasing function with non-degenerate projections. With the notations above, there are  $C_1, C_2, C_3 > 0$  such that*

$$C_1 \leq \frac{b_t}{tv_t^{2k-2} \max(1, v_t^{-k+1})} \leq C_2,$$

and

$$\begin{aligned} d_W(\tilde{F}_t, N) &\leq C_3 t^{-1/2} \max(1, v_t^{-k+1})^{1/2} \\ d_K(\tilde{F}_t, N) &\leq C_3 t^{-1/2} \max(1, v_t^{-k+1})^{1/2}. \end{aligned}$$

Concerning the bound for Kolmogorov distance, it is not formally present in the literature. It relies on the fact that in Theorem 2,  $B'(F) \leq CB(F)$  for some  $C > 0$  in the case where  $\sigma \rightarrow 1$  and  $B(F) \rightarrow 0$ . Then one can simply reproduce the proof of [16], entirely based on an upper bound for  $B(F)$ .

*Remark 9.* Theorems 4 and 5-(i) can be retrieved from this theorem by setting respectively  $v_t = 1$  or  $v_t = t$ .

*Remark 10.* If some projections do vanish, the convergence rate can be modified, and the limit might not even be gaussian, as it is the case for the degenerate geometric  $U$ -statistics of Th. 5-(ii).

*Remark 11.* Depending on the asymptotic behavior of  $v_t$ , we can identify four different regimes:

1. **Long interactions:**  $v_t \rightarrow \infty$ , CLT at speed  $t^{-1/2}$ , the first chaos  $I_{1,t}(h_{t,1})$  dominates (geometric  $U$ -statistics).
2. **Constant size interactions:**  $v_t = 1$ , CLT at speed  $t^{-1/2}$ , all chaoses have the same order of magnitude (local  $U$ -statistics).
3. **Small interactions:**  $v_t \rightarrow 0, tv_t^{-k+1} \rightarrow \infty$ , CLT at speed  $(tv_t^{-k+1})^{-1/2}$ , higher order chaoses dominate. In the case of random graphs ( $k = 2$ ), the corresponding bound in  $(tv_t)^{-1/2}$  has been obtained in [15].
4. **Rare interactions:**  $tv_t^{-k+1} \rightarrow c < \infty$ , the bound does not converge to 0. In the case  $k = 2$ , it has been shown in [15] that there is no CLT but a Poisson limit in the case  $c > 0$  (see Chapter [4] for more on Poisson limits).

## 2.2 Other limits and multi-dimensional convergence

Besides the Gaussian chaoses appearing in Theorem 5-(ii), some characterizations of non-central limits have also been derived for Poisson  $U$ -statistics.

### 2.2.1 Multidimensional convergence

We consider in this section the conjoint behavior of random variables  $F_t = (F_{1,t}, \dots, F_{k,t})$  where  $F_{m,t} = I_{q_m}(h_{m,t})$  for  $1 \leq m \leq k$ , with  $h_{m,t} \in L_s^2(\mathbb{X}^{q_m})$  for some  $q_m \geq 1$ , for  $t > 0; 1 \leq m \leq k$ .

Call  $\sigma_t^2 = \sum_{m=1}^k \text{Var}(F_{m,t})$ . Any  $L^2$  candidate for the limit of  $\sigma_t^{-1} F_t$  should have as covariance matrix

$$C_{m,n} = \lim_t \sigma_t^{-2} \mathbb{E} F_{m,t} F_{n,t}, \quad 1 \leq m, n \leq k$$

if those limits exist. In this case there is indeed asymptotic normality if all contraction norms

$$\|h_{m,t} \star_r^l h_{n,t}\|$$

go to 0 for  $r = 1, \dots, q_k$ , and every  $l = 1, \dots, r \wedge (q_k - 1)$ , under technical conditions on the kernels related to technical condition of chapter [4], see [22, Th. 5.8], [3, Th. 2.4] for details. These articles contain explicit bounds on the speed of convergence with a specific distance related to thrice differentiable functions on  $(\mathbb{R}^d)^k$ , and the convergence is stable, in the sense of [3].

If now  $F_t = (F_{1,t}, \dots, F_{k,t})$  where each  $F_{m,t}$  is a  $U$ -statistic, one can consider the random vector  $G_t$  constituted by all multiple integrals with respect to kernels from the decompositions of the  $F_{m,t}$ , as defined in (2). One can then infer conditions for asymptotic normality of  $F_t$  by applying the previous considerations to  $G_t$ .

As noted in Remark 11, some  $U$ -statistics behave asymptotically as Poisson variables. Asymptotic joint laws of  $U$ -statistics can also converge to random vectors

with marginal Poisson laws, and it can also happen that they converge to an hybrid random vector which has both Gaussian and Poisson marginals, here again the reader is referred to Chapter [4].

### 2.2.2 Gamma

Similar results to those of Section 2.1 with Gamma limits have been derived by Peccati and Thäle [21] for Poisson chaoses of even order. The distance used there is

$$d_3(U, V) = \sup_{h \in \mathcal{H}^3} |\mathbb{E}h(U) - h(V)|$$

where  $\mathcal{H}^3$  is the class of functions of class  $\mathcal{C}^3$  with all first 3 derivatives uniformly bounded by 1. We again denote by  $f \star_r^l g$  the symmetrized contraction kernels.

For  $v > 0$ , let  $F(v/2)$  be a Gamma distribution with mean and variance both equal to  $v/2$ . We introduce the centered unit variance variable  $G(v) := 2F(v/2) - v$ .

**Theorem 7.** *Let  $F = I_k(h_k)$  for some even integer  $k \geq 2$ .*

*We have*

$$d_3(I_k(h_k), G(v)) \leq D_k \max\{k! \|h_k\| - 2v; \|h_k \star_p^p h_k\|; \|h_k \star_r^l h_k\|^{1/2}; \|h_k \star_{q/2}^{q/2} h_k - c_k h_k\|\}$$

where the maximum is taken over all  $p = 1, \dots, k-1$  such that  $p \neq k/2$  and all  $(r, l)$  such that  $r \neq l$  and  $l = 0$ , or  $r \in \{1, \dots, k\}$  and  $l \in \{1, \dots, \min(r, k-1)\}$ . Also

$$c_k = \frac{4}{(q/2)! \binom{q}{q/2}^2}.$$

*Remark 12.* In the case of double integrals ( $k = 2$ ), the authors of [21] provide a 4th moment theorem, in the sense that under some technical conditions, a sequence of double stochastic integrals converge to a Gamma variable if their first moments converge to those of a Gamma variable.

*Remark 13.* This result enables to give an upper bound on the speed of convergence to the second Gaussian chaos in Theorem 5 in the case  $n_1 = 2$ , if this limit is indeed a Gamma variable.

## 3 Large deviations

There are only few investigations concerning concentration inequalities for Poisson U-statistics. Most results require an nice bound on  $\sup_{\eta \in \mathbf{N}(\mathbb{X}), z \in \mathbb{X}} D_z(F) < \infty$ . For U-statistics of order  $\geq 2$  this condition is not satisfied, even if  $f$  is bounded. For U-statistics of order 1, this holds if  $\|f\|_\infty < \infty$ . Therefore we split our investigations

into a section on U-statistics of order one and on higher order local U-statistics. We start with a general result. Throughout this section we assume that  $f \geq 0$  and  $f \neq 0$ .

### 3.1 A general LDI

In this section we sketch an approach developed in [26] leading to a general concentration inequality. For two counting measures  $\eta$  and  $\nu$  we define the difference  $\eta \setminus \nu$  by

$$\eta \setminus \nu = \sum_{x \in \mathbb{X}} (\eta(x) - \nu(x))_+ \delta_x. \quad (14)$$

For  $x \in \eta$  and  $f \in L_s^1(\mathbb{X}^k)$ , we recall that

$$U(f; \eta) = \sum_{x \in \eta} F(x; \eta) \quad \text{with} \quad F(x; \eta) = \sum_{\mathbf{x}_{k-1} \in (\eta \setminus \{x\})_{\neq}^{k-1}} f(x, \mathbf{x}_{k-1}).$$

Assume that in addition to  $\eta$  a second point set  $\zeta \in \mathbf{N}(\mathbb{X})$  is chosen. The non-negativity of  $f$  yields

$$\begin{aligned} U(f; \eta) &\leq U(f; \zeta) + k \sum_{x \in \eta} F(x; \eta) \mathbb{1}(x \notin \zeta) \\ &= U(f; \zeta) + k \int F(x; \eta) d(\eta \setminus \zeta). \end{aligned}$$

The convex distance of a finite point set  $\eta \in \mathbf{N}(\mathbb{X})$  to some  $A \subset \mathbf{N}(\mathbb{X})$  was introduced in [24], and is given by

$$d_T^\pi(\eta, A) = \max_{\|u\|_{2, \eta} \leq 1} \min_{\zeta \in A} \int u d(\eta \setminus \zeta)$$

where  $u : \mathbb{X} \rightarrow \mathbb{R}_+$  is a non-negative measurable function and  $\|u\|_{2, \eta}^2 = \int u^2 d\eta$ . To link the convex distance to the U-statistic, we insert for  $u$  the normalized function  $\|F(x; \eta)\|_{2, \eta}^{-1} F(x; \eta)$  and rewrite  $U(\eta)$  in terms of the convex distance as follows:

$$\begin{aligned} d_T^\pi(\eta, A) &\geq \min_{\zeta \in A} \int \frac{1}{\|F(x; \eta)\|_{2, \eta}} F(x; \eta) d(\eta \setminus \zeta) \\ &\geq \frac{1}{k \|F(x; \eta)\|_{2, \eta}} \min_{\zeta \in A} (U(f; \eta) - U(f; \zeta)). \end{aligned}$$

If we assume  $F(x; \eta) \leq B$ , then  $\|F(x; \eta)\|_{2, \eta}^2 \leq B \sum_{x \in \eta} F(x; \eta) = BU(f; \eta)$ , which implies

$$d_T^\pi(\eta, A) \geq \frac{1}{k\sqrt{B}} \min_{\zeta \in A} \frac{U(f; \eta) - U(f; \zeta)}{\sqrt{U(f; \eta)}} \mathbb{1}(\forall x \in \eta : F(x; \eta) \leq B). \quad (15)$$

In [24], a LDI for the convex distance was proved. For  $\eta$  a Poisson point process, and for  $A \subset \mathbf{N}(\mathbb{X})$ , we have

$$\mathbb{P}(A)\mathbb{P}(d_T^\pi(\eta, A) \geq s) \leq \exp\left(-\frac{s^2}{4}\right).$$

Precisely as in [26], this concentration inequality combined with the estimate (15) yields the following theorem.

**Theorem 8.** *Assume that  $\varepsilon(\cdot)$  and  $B \in \mathbb{R}$  satisfy  $\mathbb{P}(\exists x \in \eta : F(x; \eta) > B) \leq \varepsilon(B)$ . Let  $m$  be the median of  $U(f; \eta)$ . Then*

$$\mathbb{P}(|U(f; \eta) - m| \geq u) \leq 4 \exp\left(-\frac{u^2}{4k^2 B(u+m)}\right) + 3\varepsilon(B). \quad (16)$$

In the next sections we apply this to U-statistics of order one and to local U-statistics. In the applications, the crucial ingredient is a good estimate for  $\varepsilon(B)$ .

### 3.2 LDI for first order U-statistics

There are several concentration inequalities for integrals over Poisson point processes, i.e. U-statistics of order one,

$$U(f; \eta) = \sum_{x \in \eta} f(x) = \int f d\eta, f \geq 0$$

in which case  $D_z U = f(z)$ . Assuming that  $\|f\|_\infty = B < \infty$  we have

$$\|D_z U\|_\infty \leq B.$$

A result by Houdre and Privault [10] shows that

$$\mathbb{P}(U - \|f\|_1 \geq u) \leq \exp\left(-\frac{\|f\|_1}{\|f\|_\infty} g\left(\frac{u}{\|f\|_1}\right)\right) \quad (17)$$

where  $g(u) = (1+u)\ln(1+u) - u$ ,  $u \geq 0$  and because  $f \geq 0$  the 1-norm equals the expectation  $\mathbb{E}U(f; \eta)$ . A similar result is due to Ane and Ledoux [1]. Reynaud-Bouret [27] proves an estimate involving the 2-norm  $\|f\|_2$  instead of the 1-norm. A slightly more general estimate is given by Breton et al. [5].

We could also make use of Theorem 8 and choose  $B = \|f\|_\infty$ . This yields

$$\mathbb{P}(|U(f; \eta) - m| \geq u) \leq 4 \exp\left(-\frac{u^2}{4\|f\|_\infty(u+m)}\right), \quad (18)$$

which is a slightly weaker estimate than (17).

### 3.3 LDI for local U-statistics

In this paragraph we assume that  $\mathbb{X}$  is equipped with a distance and  $B(x, r)$  denotes the ball of radius  $r$  around  $x \in \mathbb{X}$ . If  $U$  is a local U-statistic which is concentrated on a ball of radius  $\delta_t$ , we have

$$F(x, \eta) \leq \|f\|_\infty \eta(B(x, \delta_t))^{k-1}$$

$$\begin{aligned} \mathbb{P}(\exists x : F(x; \eta) > B) &\leq \mathbb{E} \sum_{x \in \eta} \mathbb{1}(F(x; \eta) > B) \\ &\leq \int_{\mathbb{X}} \mathbb{P}(F(x; \eta) > B) d\mu_t \end{aligned}$$

and it remains to estimate

$$\mathbb{P}\left(\eta(B(x, \delta_t)) > \left(\frac{B}{\|f\|_\infty}\right)^{\frac{1}{k-1}}\right).$$

We use the Chernoff bound for the Poisson distribution, namely

$$\mathbb{P}(\eta_t(B^d(x, \delta_t)) > r) \leq \inf_{s \geq 0} e^{E(e^s - 1) - sr}, \quad (19)$$

because  $\eta(B(x, \delta_t))$  is a Poisson distributed random variable with mean

$$E(x) := \mathbb{E}\eta_t(B^d(x, \delta_t)) = \mu_t(B^d(x, \delta_t)) \leq \sup_{x \in \mathbb{X}} \mu_t(B^d(x, \delta_t)) =: E. \quad (20)$$

Because  $\inf_{s \geq 0} E(e^s - 1) - sr = r(1 - \ln(r/E)) - E$  we estimate the right hand side of (19) by  $\exp(-\frac{1}{2}r)$  for  $Ee^2 \leq r$ . This leads to

$$\mathbb{P}(\exists x : F(x; \eta) > B) \leq \mu_t(\mathbb{X}) \exp\left(-\frac{1}{2} \left(\frac{B}{\|f\|_\infty}\right)^{\frac{1}{k-1}}\right) := \varepsilon(B)$$

for  $B \geq E^{k-1} e^{2(k-1)} \|f\|_\infty$ . We set  $B = \|f\|_\infty^{\frac{1}{k}} \left(\frac{u^2}{u+m}\right)^{\frac{k-1}{k}}$  and combine this with the general Theorem 8.

**Theorem 9.** Set  $E := \sup_{x \in \mathbb{X}} \mu_t(B^d(x, \delta_t))$ . Then for  $\frac{u^2}{(u+m)} \geq E^k e^{2k} \|f\|_\infty$ ,

$$\mathbb{P}(|U(f; \eta) - m| \geq u) \leq 4\mu_t(\mathbb{X}) \exp\left(-\frac{1}{4k^2} \|f\|_\infty^{-\frac{1}{k}} \left(\frac{u^2}{u+m}\right)^{\frac{1}{k}}\right).$$

Clearly, in particular situations more careful choices of  $\varepsilon(B)$  and  $B$  lead to more precise bounds.



## 4 Applications

In this section we investigate some applications of the previous theorems in stochastic geometry. In all these cases  $\mathbb{X}$  is either a subset of  $\mathbb{R}^d$  or a subset of the affine Grassmannian  $\mathcal{A}_i^d$ , the space of all  $i$ -dimensional spaces in  $\mathbb{R}^d$ .

We state some normal approximation and concentration results which follow from the previous theorems. In many cases multi-dimensional convergence and convergence to other limit distributions can be proved in various regimes. We restrict our presentation to certain ‘simple’ cases without making any attempt for completeness. Our aim is just to indicate recent trends, we refer to further results and investigations in the literature.

### 4.1 Intersection process

Let  $\eta_t$  be a Poisson process on the space  $\mathcal{A}_i^d$  with an intensity measure of the form  $\mu(\cdot) = t\theta(\cdot)$  with  $t \in \mathbb{R}^+$  and a  $\sigma$ -finite non-atomic measure  $\theta$ . The Poisson flat process is only observed in a compact convex window  $W \subset \mathbb{R}^d$  with interior points. Thus, we can view  $\eta_t$  as a Poisson process on the set  $\mathbb{X} = [W]$  defined by

$$[W] = \left\{ h \in \mathcal{A}_i^d : h \cap W \neq \emptyset \right\}.$$

Given the hyperplane process  $\eta_t$ , we investigate the  $(d - k(d - i))$ -flats in  $W$  which occur as the intersection of  $k$  planes of  $\eta_t$ . Hence we assume  $k \leq d/(d - i)$ . In particular, we are interested in the sum of their  $j$ -th intrinsic volumes given by

$$\Phi_t = \Phi_t(W, i, k, j) = \frac{1}{k!} \sum_{(h_1, \dots, h_k) \in \eta_t^k} V_j(h_1 \cap \dots \cap h_k \cap W)$$

for  $j = 0, \dots, d - k(d - i)$ ,  $i = 0, \dots, d - 1$  and  $k = 1, \dots, \lfloor d/(d - i) \rfloor$ . For the definition of the  $j$ -th intrinsic volume  $V_j(\cdot)$  we refer to the Chapter 2 of the current book. We remark that  $V_0(K)$  is the Euler characteristic of the set  $K$ , and that  $V_n(K)$  of an  $n$ -dimensional convex set  $K$  is the Lebesgue measure  $\ell_n(K)$ . Thus  $\Phi_t(W, i, 1, 0)$  is the number of flats in  $W$  and  $\Phi_t(W, i, k, d - k(d - i))$  is the  $(d - k(d - i))$ -volume of their intersection process. To ensure that the expectations of these random variables are neither 0 nor infinite, we assume that  $0 < \theta([W]) < \infty$ , and that  $2 \leq k \leq \lfloor d/(d - i) \rfloor$  independent random hyperplanes on  $[W]$  with probability measure  $\theta(\cdot)/\theta([W])$  intersect in a  $(d - k(d - i))$ -flat almost surely and their intersection flat hits the interior of  $W$  with positive probability. For example, these conditions are satisfied if the hyperplane process is stationary and the directional distribution is not concentrated on a great subsphere.

The fact that the summands in the definition of  $\Phi_i^k$  are bounded and have a bounded support makes sure that all moment conditions are satisfied and we can apply Theorem 5:

**Theorem 10.** *Let  $N$  be a standard Gaussian random variable. Then constants  $c = c(W, i, k, j)$  exist such that*

$$\begin{aligned} d_W(\tilde{\Phi}_t, N) &\leq ct^{-1/2}, \\ d_K(\tilde{\Phi}_t, N) &\leq ct^{-1/2}, \end{aligned}$$

for  $t \geq 1$ .

Furthermore, it can be shown [25] that the asymptotic variances satisfies  $\text{Var}\Phi_t = C_\Phi t^{2k-1}(1 + o(1))$  as  $t \rightarrow \infty$  with a constant  $C_\Phi = C_\Phi(W, i, k, j)$ . The order of magnitude already follows from the first part of Theorem 5.

For more information we refer to [9] and [18]. In the second paper the Wiener-Itô chaos expansion is used to derive even multivariate central limit theorems in an increasing window for much more general functionals  $\Phi$ .

## 4.2 Flat processes

For  $i < \frac{d}{2}$  two  $i$ -dimensional planes in general position will not intersect. Thus the intersection process described in the previous section will be empty with probability one. A natural way to investigate the geometric situation in this setting is to ask for the distances between this  $i$ -dimensional planes, or more general for the so-called proximity functional. The central limit theorems described in the following fits precisely into the setting of this contribution, we refer to [30] for further results.

Let  $\eta_t$  be a Poisson process on the space  $A(d, i)$  with an intensity measure of the form  $\mu_t(\cdot) = t\theta(\cdot)$  with  $t \in \mathbb{R}^+$  and a  $\sigma$ -finite non-atomic measure  $\theta$ . The Poisson flat process is observed in a compact convex window  $W \subset \mathbb{R}^d$ . To two  $i$ -dimensional planes in general position there is a unique segment  $[x_1, x_2]$  with

$$d(h_1, h_2) = \|x_2 - x_1\| = \min_{y \in h_1, z \in h_2} \|z - y\|.$$

The midpoints  $m(h_1, h_2) = \frac{1}{2}(x_1 + x_2)$  form a point process of infinite intensity, hence we restrict this to the point process

$$\{m(h_1, h_2) : d(h_1, h_2) \leq \delta, h_1, h_2 \in \eta_\neq^2\}$$

and are interested in the number of midpoints in  $W$ .

$$\Pi_t = \Pi_t(W, \delta) = \frac{1}{2} \sum_{h_1, h_2 \in \eta_\neq^2} \mathbb{1}(d(h_1, h_2) \leq \delta, m(h_1, h_2) \in W)$$

It is not difficult to show that  $\mathbb{E}\Pi_t$  is of order  $t^2\delta^{d-2i}$ . The U-statistic  $\Pi_t$  is local on the space  $\mathcal{A}_i^d$ . Thus the following theorem due to Schulte and Thäle [30] is in spirit similar to Theorem 4.

**Theorem 11.** *Let  $N$  be a standard Gaussian random variable. Then constants  $c(W, i)$  exist such that*

$$d_K(\tilde{\Pi}_t, N) \leq c(W, i)t^{-\frac{d-i}{2}}.$$

for  $t \geq 1$ .

Moreover, Schulte and Thäle proved that the ordered distances form after suitable rescaling asymptotically an inhomogeneous Poisson point process on the positive real axis.

We add to this a concentration inequality which follows immediately from Theorem 9. Observe that  $\mu_t(\mathbb{X}) = t\theta([W])$ .

**Theorem 12.** *Denote by  $m$  the median of  $\Pi_t$ . Then*

$$\mathbb{P}(|\Pi_t - m| \geq u) \leq 4t\theta([W])\exp\left(-\frac{1}{16}\frac{u}{\sqrt{u+m}}\right)$$

for  $\frac{u}{\sqrt{u+m}} \geq e^2 t \sup_{h \in [W]} \theta(B^d(h, \delta))$ .

### 4.3 Gilbert graph

Let  $\eta_t$  be a Poisson point process on  $\mathbb{R}^d$  with an intensity-measure of the form  $\mu_t(\cdot) = t\ell_d(\cdot \cap W)$ , where  $\ell_d$  is Lebesgue measure and  $W \subset \mathbb{R}^d$  a compact convex set with  $\ell_d(W) = 1$ . Let  $(\delta_t : t > 0)$  be a sequence of positive real numbers such that  $\delta_t \rightarrow 0$ , as  $t \rightarrow \infty$ . The random geometric graph is defined by taking the points of  $\eta_t$  as vertices and by connecting two distinct points  $x, y \in \eta_t$  by an edge if and only if  $\|x - y\| \leq \delta_t$ . The resulting graph is called Gilbert graph.

There is a vast literature on the Gilbert graph and one should have a look at Penrose's seminal book [23]. More recent developments are due to Bourguin and Peccati [3], Lachièze-Rey and Peccati [15, 16] and Reitzner, Schulte and Thäle [26].

In a first step one is interested in the number of edges

$$N_t = N_t(W, \delta_t) = \frac{1}{2} \sum_{(x,y) \in \eta_{t,\neq}^2} \mathbb{1}(\|x - y\| \leq \delta_t)$$

of this random geometric graph. It is natural to consider instead of the norm functions  $\mathbb{1}(f(y-x) \leq \delta_t)$  and instead of counting more general functions  $g(y-x)$ :

$$\sum_{(x,y) \in \eta_{t,\neq}^2} \mathbb{1}(f(y-x) \leq \delta_t) g(y-x).$$

For simplicity we restrict our investigations in this survey to the number of edges  $N_t$  in the thermodynamic setting where  $t\delta_t^d$  tends to a constant as  $t \rightarrow \infty$ . Further results for other regimes, multivariate limit theorems and sharper concentration inequalities can be found in Penrose's book and the papers mentioned above.

Because of the local definition of the Gilbert graph,  $N_t$  is a local U-statistic. Theorem 6 with  $v_t = t\delta_t^d$  can be applied.

**Theorem 13.** *Let  $N$  be a standard Gaussian random variable. Then constants  $c(W)$  exist such that*

$$\begin{aligned} d_W(\tilde{N}_t, N) &\leq c(W)t^{-1/2}, \\ d_K(\tilde{N}_t, N) &\leq c(W)t^{-1/2}. \end{aligned}$$

for  $t \geq 1$ .

A concentration inequality follows immediately from Theorem 9. Observe that  $\mu_t(\mathbb{X}) = t\ell_d(W)$ .

**Theorem 14.** *Denote by  $m$  the median of  $N_t$ . Then there is a constant  $c_d$  such that*

$$\mathbb{P}(|N_t - m| \geq u) \leq 4t\ell_d(W) \exp\left(-\frac{1}{16} \frac{u}{\sqrt{u+m}}\right)$$

for  $\frac{u}{\sqrt{u+m}} \geq c_d$ .

In [26] a concentration inequality for all  $u \geq 0$  is given using a similar but more detailed approach.

#### 4.4 Random simplicial complexes

Given the Gilbert graph of a Poisson point process  $\eta_t$  we construct the Vietoris-Rips complex  $R(\delta_t)$  by calling  $F = \{x_{i_1}, \dots, x_{i_{k+1}}\}$  a  $k$ -face of  $R(\delta_t)$  if all pairs of points in  $F$  are connected by an edge in the Gilbert graph. Observe that e.g. counting the number  $N_t^{(k)}$  of  $k$ -faces is equivalent to a particular subgraph counting. By definition this is a local U-statistics given by

$$N_t^{(k)} = N_t^{(k)}(W, \delta_t) = \frac{1}{(k+1)!} \sum_{x_1, \dots, x_{k+1} \in \eta_t^{k+1}} \mathbb{1}(\|x_i - x_j\| \leq \delta_t, \forall 1 \leq i, j \leq k+1).$$

Central limit theorems and a concentration inequality follow immediately from the results for local U-statistics. We restrict our statements again to the thermodynamic case where  $t\delta_t^d$  tends to a constant as  $t \rightarrow \infty$ . Results for other regimes can be found e.g. in Penrose's book. Because of the local definition of the Gilbert graph,  $N_t^{(k)}$  is a local U-statistic. Theorem 6 with  $v_t = t\delta_t^d$  can be applied.

**Theorem 15.** *Let  $N$  be a standard Gaussian random variable. Then constants  $c(W)$  exist such that*

$$\begin{aligned} d_W(\tilde{N}_t^{(k)}, N) &\leq c(W)t^{-1/2}, \\ d_K(\tilde{N}_t^{(k)}, N) &\leq c(W)t^{-1/2}. \end{aligned}$$

for  $t \geq 1$ .

A concentration inequality follows immediately from Theorem 9. Observe that  $\mu_t(\mathbb{X}) = t\theta([W])$ .

**Theorem 16.** *Denote by  $m$  the median of  $N_t$ . Then*

$$\mathbb{P}(|\Pi_t - m| \geq u) \leq 4t\ell_d(W) \exp\left(-\frac{1}{4(k+1)^2} \frac{u^{\frac{2}{k}}}{(u+m)^{\frac{1}{k}}}\right)$$

for  $\frac{u^2}{u+m} \geq c_d$ .

Much deeper results concerning the topology of random simplicial complexes are contained in [6, 12] and [14]. We refer the interested reader to the recent survey article by Kahle [13]

#### 4.5 Sylvester's constant

Again we assume that the Poisson point process  $\eta$  has an intensity-measure of the form  $\mu_t(\cdot) = t\ell_d(\cdot \cap W)$ , where  $\ell_d$  is Lebesgue measure and  $W \subset \mathbb{R}^d$  a compact convex set with  $\ell_d(W) = 1$ .

As a last example of a U-statistic we consider the following functional related to Sylvester's problem. Originally raised with  $k = 4$  in 1864, Sylvester's original problem asks for the distribution of the number of vertices of the convex hull of four random points. Put

$$N_t = N_t(W, k) = \sum_{(x_1, \dots, x_k) \in \eta_{\neq}^k} \mathbb{1}(x_1, \dots, x_k \text{ are vertices of } \text{conv}(x_1, \dots, x_k)),$$

which counts the number of  $k$ -tuples of the process such that every point is a vertex of the convex hull, i.e., the number of  $k$ -tuples in convex position.

The expected value of  $U$  is then given by

$$\mathbb{E}N_t = t^k \mathbb{P}(X_1, \dots, X_k \text{ are vertices of } \text{conv}(X_1, \dots, X_k)) = t^k p(W, k),$$

where  $X_1, \dots, X_k$  are independent random points chosen according to the uniform distribution on  $W$ .

The question to determine the probability  $p(W, k)$  that  $k$  random points in a convex set  $W$  are in convex position has a long history, see e.g. the more recent de-

velopment by Bárány [2]. In our setting, the function  $t^{-k}N_t$  is an estimator for the probability  $p(W, k)$  and we are interested in its distributional properties.

The asymptotic behaviour of  $\text{Var}(N_t)$  is of order  $t^{2k-1}$ . Together with Theorem 5, we immediately get the following result showing that the estimator  $H$  is asymptotically Gaussian:

**Theorem 17.** *Let  $N$  be a standard Gaussian random variable. Then there exists a constant  $c(W, k)$  such that*

$$d_W(\tilde{N}_t, N) \leq c(W, k)t^{-\frac{1}{2}}.$$

## References